A HIPAA-Compliant Document Redaction Tool Using Natural Language Processing for Sensitive Information Protection

Donald C. Cooper†
Medical & Data Science Division
Ramos Law
Northglenn, CO USA
DrDon@Ramoslaw.com

Channing Cooper
Medical & Data Science Division
Ramos Law
Northglenn, CO USA
C.Cooper@Ramoslaw.com

ABSTRACT

The integration of sensitive medical and legal documents into AI-driven workflows requires robust de-identification to comply with privacy regulations like HIPAA while preserving data utility. This paper presents a scalable Natural Language Processing (NLP) pipeline designed to automate the removal of all 18 HIPAA-defined protected health identifiers (PHI) from PDF-based medical and legal records. The system converts PDF documents into plain text, applies a combination of rule-based pattern matching, named entity recognition (NER), and context-aware filtering to detect and redact sensitive information—including names, dates, geographic identifiers, and medical record numbers. The sanitized output is validated against HIPAA's privacy criteria, ensuring compliance before being securely passed to large language models (LLMs) via API for downstream tasks such as summarization, trend analysis, or legal case abstraction. We demonstrate the pipeline's efficacy on a corpus of real-world medical reports and legal documents, achieving near-complete PHI removal while retaining critical contextual information for AI analysis. The tool addresses a key bottleneck in healthcare and legal AI adoption by enabling institutions to leverage LLMs without compromising patient or client confidentiality. Its open-source implementation offers a practical, auditable solution for organizations seeking to balance innovation with regulatory compliance.

CCS CONCEPTS

• Security and Privacy

KEYWORDS

HIPAA Compliance, De-identification, NLP Pipeline, Medical-Legal Documents, LLM Integration

ACM Reference format:

Donald Cooper, Channing Cooper 2025. A HIPAA-Compliant Document Redaction Tool Using Natural Language Processing for Sensitive Information Protection

1 INTRODUCTION

With the recent rise in artificial intelligence (AI)-based tools, the protection of sensitive information has become a paramount concern, especially in industries such as law and healthcare, where the handling of Personally Identifiable Information (PII) and Protected Health Information (PHI) is governed by strict regulations like HIPAA. Unauthorized disclosure of such information can lead to severe legal consequences, financial penalties, and loss of trust. Manual redaction of sensitive information from documents is time-consuming, error-prone, and often inadequate for large-scale operations. Automated redaction tools offer a promising solution, but they must balance accuracy, efficiency, and compliance with legal standards.

This paper introduces a HIPAA-compliant document redaction tool that combines Optical Character Recognition (OCR) and Natural Language Processing (NLP) to automatically detect and redact sensitive information from documents. The tool is designed to handle various document formats, including PDFs and images, and allows users to specify custom fields for redaction. By automating the redaction process, the tool reduces the risk of human error and ensures compliance with HIPAA regulations, making it a valuable asset for healthcare providers, legal professionals, and organizations handling sensitive data.

2 METHODS

2.1 System Architecture

The redaction tool is built as a web application using the python programming language and the Flask framework, which provides a lightweight and flexible platform for handling document uploads, processing, and redaction. The system architecture consists of the following key components:

1. **User Interface (UI):** The front-end interface allows users to upload documents, specify sensitive information fields, and view redacted outputs. The UI is designed to be intuitive, with clear instructions and feedback mechanisms to guide users through the redaction process.

2. **Document Processing Engine:** The core of the system is the document processing engine, which handles the extraction of text from uploaded documents. For PDF files, the tool uses the PyMuPDF library to extract text,

while OCR (via Tesseract) is employed for image-based documents. The engine ensures that text is accurately extracted, even from scanned documents or low-quality images.

3. **Redaction Module:** The redaction module uses a combination of regular expressions, NLP, and fuzzy matching to identify and redact sensitive information. The module supports a wide range of PHI fields, including names, dates of birth, Social Security numbers, and medical record numbers. Users can also specify custom fields for redaction, making the tool adaptable to different use cases.

4. **Metrics and Reporting:** After redaction, the tool generates detailed metrics, including the number of redactions, the percentage of text modified, and the types of fields redacted. These metrics provide users with insights into the redaction process and help ensure that all sensitive information has been adequately removed.

5. **Access Control:** To ensure that only authorized users can access the tool, an access control mechanism is implemented. Users must enter a valid access code, which is validated against a database of authorized codes. This feature adds an additional layer of security, ensuring that the tool is used only by authorized personnel.

2.2 Redaction Algorithm

The redaction algorithm is designed to handle various types of sensitive information, including structured data (e.g., dates, phone numbers) and unstructured data (e.g., names, addresses). The algorithm employs the following steps:

1. **Text Extraction:** The tool first extracts text from the uploaded document using OCR for images or PDF text extraction for PDF files. If text extraction fails or yields insufficient results, OCR is applied to the entire document.

2. **Field-Specific Redaction:** The tool uses regular expressions and NLP to identify and redact specific fields of sensitive information. For example, dates of birth are redacted using a combination of date format patterns and fuzzy matching, while names are redacted using NLP-based entity recognition and fuzzy string matching.

3. **Fuzzy Matching:** To account for variations in how sensitive information is presented (e.g., misspellings, abbreviations), the tool employs fuzzy matching techniques. This ensures that even partially matching strings are redacted, reducing the risk of information leakage.

4. **Redaction Application:** Once sensitive information is identified, it is replaced with the "[REDACTED]" tag. The tool preserves the original document structure, ensuring that the redacted document remains readable and usable.

5. **Output Generation:** The redacted text is saved as a new file, and users are provided with a download link. The tool also generates a report detailing the redaction process, including the number of redactions and the types of fields redacted.

2.3 Evaluation Metrics

The effectiveness of the redaction tool is evaluated using several metrics, including:

- **Redaction Accuracy:** The percentage of sensitive information correctly identified and redacted.
- **False Positives:** The number of non-sensitive items incorrectly redacted.
- **Processing Time:** The time taken to process and redact a document.
- **User Satisfaction:** Feedback from users on the tool's ease of use and effectiveness.

3. RESULTS

The redaction tool was tested on a dataset of 100 documents, including PDFs and images, containing various types of sensitive information. The results demonstrate the tool's effectiveness in redacting sensitive information while maintaining document integrity:

- **Redaction Accuracy:** The tool achieved an accuracy of 98% in identifying and redacting sensitive information, with minimal false positives.
- **Processing Time:** The average processing time for a 10-page document was 12.3 seconds, making the tool suitable for real-time use.
- **User Satisfaction:** Users reported high satisfaction with the tool's ease of use and the clarity of the redaction reports.

4. DISCUSSION

The HIPAA-compliant document redaction tool presented in this paper offers a robust solution for protecting sensitive information in digital documents. By combining OCR, NLP, and fuzzy matching, the tool achieves high accuracy in identifying and redacting sensitive information, while maintaining document

readability. The tool's user-friendly interface and detailed reporting make it accessible to a wide range of users, from healthcare providers to legal professionals.

One of the key strengths of the tool is its adaptability. Users can specify custom fields for redaction, making the tool suitable for a variety of use cases beyond healthcare, such as legal document redaction or financial data protection. Additionally, the tool's access control mechanism ensures that it is used only by authorized personnel, adding an extra layer of security.

However, there are some limitations to the current implementation. The tool relies heavily on the quality of the input documents, particularly for OCR-based text extraction. Low-quality scans or handwritten text may result in reduced accuracy. Future work could explore the integration of advanced machine learning models to improve text extraction and redaction accuracy, particularly for challenging document types.

Acknowledgments:
The authors would like to thank the developers of the open-source libraries used in this project, including Flask, Tesseract, and spaCy, for their contributions to the field of document processing and NLP.

REFERENCES

[1] Health Insurance Portability and Accountability Act (HIPAA), 1996.